1V-02

Character-Level CNNを用いた日本語評判分析

田村 匠† 丸山 真佐夫†

†木更津工業高等専門学校 情報工学科

1 はじめに

画像処理分野で用いられるConvolutional Neural Networks (CNN) を自然言語処理に応用したCharacter-Level CNN (CLCNN) が近年注目されている．多くの既存手法では外部の辞書データやライブラリを用いて，形態素解析をはじめとする前処理を行うことが多いが，CLCNNにはそのような前処理が不要であるという特徴がある．また，英語においてCLCNNは，顧客からのレビューを扱う評判分析に効果的であることがわかっている [4]．評判分析は実社会での応用が特に盛んなタスクであるが，日本語におけるCLCNNを用いた評判分析の効果はデータセットの不足などもあり十分に検証されていない．そこで，本研究ではCLCNNを用いて日本語の評判分析を行い，既存手法と比較することで，その性能を評価した．

2 Character-Level CNNと関連先行研究

2015年にZhangらが発表したCLCNN[4]は，文章を単語単位ではなく文字単位でone-hotベクトルに変換し，それらのベクトルを各行として並べた2次元データをCNNモデルに入力する手法である．Zhangらは CLCNNが英語の評判分析で効果的であることを示したが，日本語の文字種は膨大であり，同様の手法を用いることはできない．そこで，Satoらは文字埋め込み (Character Embedding) によってone-hotベクトルを任意次元の密ベクトルとすることで，日本語にCLCNNを応用し，ECサイトのレビューを分類する評判分析を行った [3]．また，宗里らはZhangらやSatoらと構造の異なる新しいモデルによって記事からの新聞社推定を行い [5]，宮崎らは宗里らと同構造のモデルを用いて病気に関わるツイートの分類を行っている [7]．

3 実験手法

CLCNNを用いた日本語評判分析の性能を評価するために，CLCNNと比較既存手法のそれぞれにおいて2つの日本語評判分析データセットを用いて，あるレビューがポジティブ (*pos*) なのか，ネガティブ (*neg*) なのかを判定する2値分類 (肯否判定) を行った．実験には4-fold Cross Validationを用い，その平均正答率を評価対象とした．

3.1 CLCNN

先行研究 [5][7] を参考にほぼ同構造のCLCNNモデルを構築した．データセットに含まれる140文字以内のレビューの各文字をUnicode符号位置に変換し，文字埋め込みによって128次元の密ベクトルに変換することで日本語CLCNNを行った．出力層の活性化関数にはsoftmax関数を使用し，モデルが *pos* と *neg* である確率をそれぞれ出力するようにした．出力層以外の活性化関数にはReLUを用い，誤差関数には交差エントロピーを，確率的最急降下法にはAdamを利用し，過学習を防ぐため学習データの1割をValidation Dataとし Early Stoppingによって学習回数を動的に決定した．

3.2 比較既存手法

CLCNNを用いた日本語評判分析の性能を評価するために既存手法として3つの手法で比較を行った．

oseti[1]

osetiは形態素解析を行った上で，日本語極性評価辞書を利用して肯否判定を行う感情分析ライブラリである．学習を必要としない手法であるため，4つのfoldに対する正答率の平均を評価値とした．

Bag-of-Words (BoW)

単語の出現頻度などを特徴量とするBoWは，教師あり学習を利用した既存手法として一般的である．そこで，学習データからTF-IDF特徴量を算出し，多項分布を用いたナイーブベイズ分類器を用いて肯否判定を行った．単語分割のための形態素解析にはMeCabを利用した．

Bag-of-Ngrams (BoN)

文字Ngramを用いて単語分割を行うことで，CLCNNと同じように形態素解析を行わずに，BoWを用いることができる．N の値 (複数も可) はハイパーパラメータとなるが，今回の実験では最も良かった結果を評価値とした．

3.3 データセット

実験に利用するデータセットとして，Twitter日本語評判データセットとJapanese Realistic Textual Entailment Corpus (jrte-corpus)[2][6]を利用した．前者は携帯電話などに関するツイートのIDと肯否ラベルで構成されたデータセットで，取得できたツイートのうち *pos* と *neg* をそれぞれ9,036件無作為抽出した．後者は旅行サイト「じゃらん」の短いクチコミデータに対して肯否ラベルがつけられたデータセットで，*pos* と *neg* をそれぞれ818件無作為抽出した．

3.4 前処理と辞書選択

自然言語処理では，正規化や記号除去といった前処理の有無によって結果が大きく変化する可能性がある．そこで，前処理ライブラリneologdnを用いた**正規化**と，正規化に加えて記号や絵文字を除去する**記号除去**という，2つの前処理をデータに施し，**無処理**の場合とそれぞれ比較した．また

Sentiment analysis using Character-Level CNN in Japanese

Takumi Tamura†, Masao Maruyama†

†Department of Information and Computer Engineering, National Institute of Technology, Kisarazu College

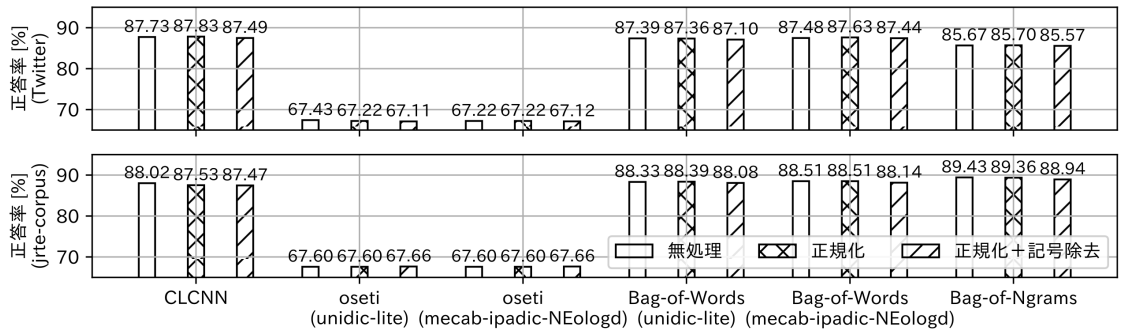

図 1　各手法・各前処理・各データセットにおける平均正答率

oseti と BoW では，形態素解析を行う際に辞書が必要になるが，辞書の語彙も結果に大きく影響する可能性がある．今回の実験では，小規模な辞書である unidic-lite と，インターネット上の新語などを収録している大規模な辞書である mecab-ipadic-NEologd のそれぞれを用いて比較した．

4　実験結果

各手法・各前処理・各データセットにおける平均正答率を図 1 に示した．

極性評価辞書を用いる oseti はどの場合においても正答率が 67%台であった．他の機械学習を用いた手法と比較すると，CLCNN はそれらの手法とほぼ同等の性能である．Twitter 日本語評判データセットでは CLCNN が最も性能が良かったが，jrte-corpus では BoN が最も性能が良かった．CLCNN の学習曲線を分析すると，CLCNN モデルは high variance の傾向が見られ，学習データを増やすことで正答率の改善が期待できる．とくに jrte-corpus はデータ数が Twitter 日本語評判データセットの 1/10 未満であり，学習データ不足の傾向が強い．

正規化はいくつかの場合において正答率を向上させたが，下落させることもあった．正規化は常に最適な手法ではないといえるが，全体として正規化の与える影響は軽微であった．また記号除去は，ほとんどの場合において，正規化のみの結果と比較して正答率の低下が見られた．この事実は，肯否判定において，記号や絵文字が一定の役割を果たしていることを示唆する．また，BoW における辞書の選択では，若干ながら大規模辞書の方が精度が良いことがわかる．

jrte-corpus で最も性能が良かった BoN では，ハイパーパラメータ N が存在する．Twitter 日本語評判データセットでは $N = 3$ の場合が最も性能が良かったのに対して，同様に jrte-corpus で実験を行うと正答率は 85%台に下落した．一方で，CLCNN では両データセットにおいて全く同一のハイパーパラメータを適用している．今回の研究で利用した宗里らのモデル [5] は内部で複数の Ngram を並列して行うような複雑な構造を持っており，これがハイパーパラメータに影響されづらい理由であると考えられる．

総じて，CLCNN は形態素解析を行う BoW とほぼ同等であり，BoN よりもハイパーパラメータやデータセットに影響されづらい．それ故にモデルは複雑であり，学習データが少ない場合は BoW，BoN と比べて不利である．

5　おわりに

実験の結果から，日本語の評判分析において CLCNN は，形態素解析などを利用する既存手法とほぼ同等の性能であり，ハイパーパラメータやデータセットに影響されづらい汎用的な手法であることがわかった．CLCNN は既存手法と比較すると，学習や検証に十分な時間や学習データ量が必要なことなど欠点も存在するが，評判分析の新しい手法として様々な用途に応用が期待される．今回の研究結果をもとに，CLCNN の改善や評判分析以外の他の分野への応用に取り組んでいきたい．

謝辞

本研究では，鈴木優氏の “Twitter 日本語評判データセット”(http://www.db.info.gifu-u.ac.jp/sentiment_analysis) と，株式会社リクルートが提供する “Japanese Realistic Textual Entailment Corpus” (https://github.com/megagonlabs/jrte-corpus) を利用しました．

参考文献

[1] ikegami-yukino/oseti: Dictionary based sentiment analysis for japanese. https://github.com/ikegami-yukino/oseti.

[2] Yuta Hayashibe. Japanese realistic textual entailment corpus. In *Proc. LREC*, 2020.

[3] Minato Sato, Ryohei Orihara, Yuichi Sei, Yasuyuki Tahara, and Akihiko Ohsuga. Japanese text classification by character-level deep convnets and transfer learning. In *Proc. ICAART*, 2017.

[4] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Proc. NIPS*, 2015.

[5] 宗里駿, 小谷龍ノ介, 彌冨仁. Character-level Convolutional Neural Networks を用いた新聞社間の記事の違いの解析の試み. 言語処理学会第 24 回年次大会論文集, 2018.

[6] 林部祐太. 知識の整理のための根拠付き自然文間含意関係コーパスの構築. 言語処理学会第 26 回年次大会論文集, 2020.

[7] 宮崎和光, 井田正明. テキスト分析における Character-level CNN の性能評価-NTCIR-13MedWeb タスクを題材として-. 第 48 回知能システムシンポジウム講演資料, 2021.